Title

The early embryonic transcriptome of a Hawaiian *Drosophila* picture-wing fly shows evidence of altered gene expression and novel gene evolution

Running Title

Transcriptome Hawaiian *Drosophila* embryo

Authors

Madeline Chenevert¹, Bronwyn Miller¹, Ahmad Karkoutli¹, Anna Rusnak¹, Susan Lott² and Joel Atallah¹

Affiliations when the work was carried out

¹Department of Biological Sciences, University of New Orleans, New Orleans, LA, USA 70148

²Department of Evolution & Ecology, University of California - Davis, Davis, CA, USA

Current Affiliations

Madeline Chenevert: Hayward Genetics Center, Tulane University School of Medicine, 1430 Tulane Avenue, SL-31, New Orleans, LA 70112

Anna Rusnak: Center for Biomedical Engineering, Brown University, Box A-2, Arnold Lab, 91 Waterman Street, Suite 211, Providence, RI 02912

Ahmad Karkoutli: LSUHSC School of Medicine, 1901 Perdido St, New Orleans, LA 70112

ABSTRACT

A massive adaptive radiation on the Hawaiian archipelago has produced approximately one quarter of the fly species in the family Drosophilidae. The Hawaiian *Drosophila* clade has long been recognized as a model system for the study of both the ecology of island endemics and the evolution of developmental mechanisms, but relatively few genomic and transcriptomic datasets are available for this group. We present here a differential expression analysis of the transcriptional profiles of two highly conserved embryonic stages in the Hawaiian picture-wing

fly *Drosophila grimshawi*. When we compared our results to previously published datasets across the family Drosophilidae, we identified cases of both gains and losses of gene representation in *D. grimshawi*, including an apparent delay in Hox gene activation. We also found high expression of unannotated genes. Most transcripts of unannotated genes with open reading frames do not have homologs in non-Hawaiian *Drosophila* species, although the vast majority have sequence matches in other genomes of the Hawaiian picture-wing flies. Some of these genes may have arisen from non-coding sequence in the ancestor of Hawaiian flies or during the evolution of the clade. Our results suggests that both the modified use of ancestral genes and the evolution of new ones may occur in rapid radiations.

KEYWORDS

Hawaiian *Drosophila*, embryo, transcriptomics, maternal-to-zygotic-transition, Hox genes, de novo genes

RESEARCH HIGHLIGHTS

The early embryonic transcriptome of the Hawaiian fly *Drosophila grimshawi* shows a loss of expression of conserved Stage 5 genes, including the Hox genes

The de novo evolution of embryonically expressed genes may be occurring in the Hawaiian *Drosophila* lineage

AUTHORS' STATEMENT

This paper is not being considered for publication elsewhere. This study formed part of Madeline Chenevert's M.S. thesis.

INTRODUCTION

It is estimated that 1,000 of the world's 4,000 *Drosophilid* fly species are endemic to the Hawaiian archipelago (O'Grady & DeSalle, 2018). Comprising the largest, and arguably the most diverse, radiation within the family, Hawaiian *Drosophila* flies (Figure 1) show extensive phenotypic variation, with morphological modifications at every stage of the life cycle (Craddock et al., 2018; Sarikaya et al., 2019). Flies of the iconic Hawaiian picture-wing clade (Magnacca & Price, 2015), for example, show highly varied wing patterns that may be sexually monomorphic or dimorphic (Edwards et al., 2007). Other clades are named after characteristic modifications to the tarsus (Lapoint et al., 2009, 2014) or mouth parts (Magnacca & Grady, 2009), both specific to male flies. In addition to changes to the adult body plan, the eggs display a vast range of variation in size, structure, and filament number (Kambysellis et al., 1980; Kambysellis & Heed, 1971). This phenotypic diversity raises the question of whether there is similar variation in gene expression, and whether the evolution of novel genes (Van Oss & Carvunis, 2019), or the re-use of ancestral ones (Carroll, 2008), underlies the diversification of the clade.

Transcriptional profiling of Hawaiian flies, however, has been limited. While recent studies have carried out functional genomics in adult stages of picture-wing flies (Kang et al., 2016), gene expression levels in the embryo have not previously been analyzed. Hawaiian flies are notoriously difficult to culture (Montgomery, 1975), and the life cycle is extended relative to common laboratory species such as *Drosophila melanogaster*, complicating both the collection of samples and the interpretation of results from timed specimens.

Early embryogenesis, which constitutes the opening act in an organism's life (Tadros & Lipshitz, 2009), is the basis for all subsequent development, and perhaps the most critical period to first consider when analyzing a species with a dearth of transcriptomic data that is difficult to culture. Across *Drosophilidae*, early embryos pass through comparable developmental stages (Kuntz & Eisen, 2014). We have previously shown (Atallah & Lott, 2018) that gene transcript levels are highly concordant across species at two early embryonic stages, stage 2 (St2) and late stage 5 (St5) (Figure 2). The first of these stages precedes the maternal-zygotic transition, while at stage 5 zygotic expression is underway and many maternal transcripts have been degraded (Figure 2). While most genes are represented at both stages, those that are St5-only (zygotically transcribed with no maternal contribution) are strongly conserved across large phylogenetic distances (Atallah & Lott, 2018).

Here, we present results of a single-embryo RNA-Seq analysis of these critical stages in the Hawaiian picture-wing fly *D. grimshawi*. We compare the findings to our previously published data on other *Drosophila* flies. We find that *D. grimshawi* shows extensive loss of gene transcript representation at zygotic genome activation (stage 5) relative to non-Hawaiian species. Notably, in sharp contrast to outgroup clades, we find no evidence that the Hox genes, the downstream component of the antero-posterior segmentation cascade, are activated by stage 5. We further find instances of early embryonic mRNA representation of gene orthologues in this picture-wing fly which are not seen in other species until later stages in development. Finally, we conduct an analysis of unannotated genes in *D. grimshawi* that are represented in the early embryo, and provide evidence that many of them are taxonomically restricted. Some of these genes appear to have been generated de novo from non-coding

sequence, either in the Hawaiian *Drosophila* proper or in the ancestor of Hawaiian *Drosophila* and its sister-clade, *Scaptomyza* (Figure 1), which is believed to have originated in Hawaii before diversifying globally (Lapoint et al., 2013; O’Grady & DeSalle, 2008).

MATERIALS AND METHODS

Drosophila grimshawi husbandry

Upon receiving the *D. grimshawi* stock (Stock Number 15287-2541.00) from the San Diego *Drosophila* Species Stock center (currently the National *Drosophila* Species Stock Center at Cornell University, <https://www.drosophilaspecies.com/>), the flies were maintained in accordance with a modified version of stock center protocols. The line we used was a replacement donated by Dr. Ken Kaneshiro from the same stock that was used in sequencing the *D. grimshawi* genome for the 12 *Drosophila* Genomes Project (Clark et al., 2007). The flies were kept in vials of Wheeler-Clayton food (Wheeler & Clayton, 1965). To counter the high humidity of New Orleans, which could have caused the flies to stick to the food, Wheeler-Clayton food was prepared using the maximum recommended amount of agar (14.5 g per liter of water). The cooked potato media was prepared with 456 g of filtered water for every 100 g of powdered mix to ensure a consistency that prevented food from falling during transfers. High doses of ethanol and propionic acid (6.5 ml per liter of water) were also used in the top layer recipe to prevent mold growth.

Vials were papered with Kimwipes, and flies were transferred to new vials twice per week. We found that maintaining stocks at a density of 12-16 flies per vial maximized fecundity while reducing death due to overcrowding. Vials containing third instar larvae were transferred to

jars to pupate. The jars contained equal volumes of oolite and aragonite sand, approximately 2 cm deep, and were covered with two layers of heavy-duty paper towels. The sand was moistened with water from a spray bottle once or twice per week, with the aim of keeping the sand and towels damp while avoiding puddles. Excess water was absorbed with cellulose acetate plugs (“flugs”, Genesee Scientific), which were then disposed of.

After the first eclosions, the vials with larvae were moved to a new sand jar, and a small petri dish containing the top layer of the Wheeler-Clayton food was added to the jar to feed new flies as they emerged. Initially, we rapidly transferred the recently-eclosed flies to vials to expand our population, but we observed that they often became stuck in the paper or condensation due to the high ambient humidity in New Orleans. To avoid losing adults before they reached sexual maturity, the flies were maintained in the jar for two weeks, then moved to large egg collection cages. The cages were placed atop petri dishes containing a base of apple-agar media, combined with 2 mL of melted Wheeler-Clayton food, providing nutrient-rich media to encourage egg-laying. These Wheeler-Clayton pellets, containing the eggs, were transferred from the apple media with a spatula, washed with water or a 50% solution of apple cider vinegar to inhibit mold, cut into multiple pieces with a spatula and mixed into vials of Wheeler-Clayton food. Most vials prepared in this manner produced dozens of larvae and little to no mold.

Egg collection

Single embryo collection and staging were carried out as described previously (Atallah & Lott, 2018), with a few modifications. Briefly, flies were maintained at room temperature in an

egg collection cage lined with a petri dish filled with apple-agar media, with 2 ml of a paste of live yeast in the center of the plate. Since stage 2 embryos can be confused with later stages, the cages were timed to ensure all eggs were laid within one hour. The flies were maintained on live yeast for an hour, after which the plate was discarded and replaced with a new one. Stage 5 embryos were collected from these plates, and Stage 2 embryos were only collected after plates were replaced once more.

Eggs were dechorionated with a 50% bleach solution. Unlike smaller eggs of non-picture wing species, which could be dechorionated after less than two minutes in bleach, dechorionation of *D. grimshawi* eggs could take upwards of three minutes. The process was usually halted when the filaments were completely dissolved and no longer visible. Eggs were rinsed with deionized water for at least 30 seconds. The dechorionated eggs were then transferred to a drop of halocarbon oil (Sigma) on a microscope slide. The slide was placed on a dissection scope over refracted light. Individual embryos were selected using a set of morphological characteristics (Bownes, 1975). St2 embryos were identified by the empty poles on each end of the egg and their lack of visible cell membranes and evagination. St5 embryos were collected at the end of stage 5, after cellularization and before gastrulation. Traits required for St5 embryos were 2 layers of well-defined cell membranes and formation of the pole cells. Embryos showing damage or any invagination were rejected.

Embryo processing and RNA extraction

A 1.5 mL microcentrifuge tube was filled with 800 µl of Trizol reagent (Ambion) and labelled with the sample number. Each embryo was placed on the corner of a clean microscope

cover slip. Using a fine paintbrush, the embryo was transferred to a new location on the slide to remove excess halocarbon oil. After repeating this process two or three times, the embryo was positioned in the center of the slide for lysing. 3 µl of Trizol were pipetted from the labelled 800-µl aliquot and placed in a droplet on top of the embryo. The embryo was then lysed with a sterile, 30-gauge medical lancet (Reli-On Ultra-Thin). Embryos were left to dissolve in the Trizol for 5 minutes. After the embryo had completely dissolved, an additional 3 µl of Trizol from the tube were added to the slide, and all 6 µl were pipetted and transferred to the tube containing the remaining 794 µl, which was mixed by pipetting multiple times. 6 microliters were again removed from the same tube and pipetted onto the place of the embryo, re-collected and returned to the aliquot. This rinsing step was repeated twice to ensure thorough collection of all embryonic tissues. Labelled samples were stored at -80°C until RNA isolation.

RNA was extracted per the manufacturer's instructions using a Trizol phenol-chloroform extraction (Invitrogen) in which the method was modified to accommodate an initial volume of 800 µl of Trizol. During RNA precipitation, 10 µL of 20 µg/µl glycogen (Invitrogen™ UltraPure™ Glycogen) was used, and the samples were spun in an Eppendorf 5424 R centrifuge at 21130 RCF (maximum speed) for a full 60 minutes. If a sample showed no visible pellet at this point, it was spun for another 30 minutes. After resuspension in UltraPure water, 15.5 µl were kept for processing, and 3 aliquots of 1.5 µl each were taken for quality analysis.

RNA Quality analysis

RNA integrity was assessed on an Experion bioanalyzer using a Bio-Rad Experion RNA High-Sensitivity kit or an Agilent 2100 bioanalyzer using a Bioanalyzer RNA 6000 Pico assay

(Agilent). RNA concentrations were measured using a Qubit 4 fluorometer with a Qubit high-sensitivity RNA assay.

cDNA libraries

To remove any DNA contamination before library construction, a Turbo DNA-Free kit was used according to the manufacturer's directions before immediately continuing to library construction using the NEBNext Ultra RNA Library Prep kit for Illumina. 15 PCR cycles were used in the final enrichment step of library generation. cDNA quality was assessed using an Agilent 2100 Bioanalyzer with a High-Sensitivity DNA Kit. The samples showed main peaks around 300 base pairs, as expected.

Sequencing

The four highest quality libraries from each stage were sent to Novogene for sequencing. Only libraries from samples showing minimal RNA degradation were selected. In the final analysis, 3 libraries from each stage were used. One library from stage 2 was not used due to low overall read mapping rate (less than 65%) while another library from stage 5 appeared to have been mislabeled. The files containing the reads are publicly available at NCBI's Sequence Read Archive (BioProject Accession Number PRJNA771180).

Transcriptome mapping, assembly and differential expression

Bioinformatic analysis was carried out as described previously (Atallah & Lott, 2018). Briefly, adapters were removed using Cutadapt (Martin, 2011) and reads were trimmed and filtered for quality. We used the Tuxedo suite (Trapnell et al., 2012) for transcriptomic analysis.

Although much faster alignment-free methods have become more popular in recent years, we chose the Tuxedo suite both to allow the identification of novel isoforms and genes and for straightforward comparison with our previous analyses of equivalent stages in non-Hawaiian species. We aligned the reads, using Tophat2 (D. Kim et al., 2013), to the GCF_000005155.2_dgri_caf1 *D. grimshawi* NCBI reference genome. Cufflinks (Trapnell et al., 2013) was used with the -N upper-quartile normalization option. Gene expression levels were determined by combining the expression of all gene isoforms. Differential expression was determined using CuffDiff. *D. melanogaster* orthologs were assigned using the Flybase (Larkin et al., 2021) orthology table. An FPKM threshold of 1 was used, as employed previously (Atallah & Lott, 2018).

Gene Ontology (GO) enrichment analysis

GO enrichment of genes with identified one-to-one *D. melanogaster* orthologs was assessed using DAVID (Huang et al., 2009a, 2009b). In cases where Cufflinks combined two annotated genes, both genes were considered differentially expressed in the GO analysis. The background list in all cases was the set of all genes represented in the *D. grimshawi* embryo, unless otherwise specified. The R package GOplot (Walter et al., 2015) was used to generate graphical representations of gene enrichment.

Unannotated gene analysis

Transcripts of genes that were identified by Cufflinks but had no annotation in the NCBI genome were searched for open reading frames using TransDecoder (Haas et al., 2013). Those with complete open reading frames of at least 50 amino acids were then compared to both the

sequenced genomes of 11 other *Drosophila* species and previously generated embryonic transcriptome assemblies of the same species (Atallah & Lott, 2018; Combs & Eisen, 2013; Lott et al., 2014; Paris et al., 2015). For cases where no ortholog was identified using BLASTP (Altschul et al., 1990), we used TBLASTN to search for matches in the other *Drosophila* genomes, then analyzed these matches for ORFs using TransDecoder. Alignments were generated using CLUSTALW (Thompson et al., 2003).

RESULTS

Gene representation in the early *D. grimshawi* embryo

We found a total of 18,617 transcripts to be represented in the early embryo above our established threshold (see Materials and Methods). These transcripts belonged to 8,233 genes. Genes with one-to-one orthologs in *D. melanogaster* or other species are shown in Table S1. Using insights obtained from previous analyses, we identified four classes of genes (Figure 2), discussed below. We used the R package GOPlot (Walter et al., 2015) to graphically represent the results of our GO analysis as concentric circular plots (Figures 3 and 4).

Class 1: Stage 2-only

Our first group of genes have maternally supplied transcripts that are completely degraded by stage 5. They form the numerically smallest class (241 genes), and an analysis using DAVID did not show significant enrichment for any GO terms when compared to the full set of embryonically represented genes (Table S2). In other words, genes with transcripts that are completely degraded in early embryogenesis appear to be randomly distributed across functional categories and cellular components. We have previously shown that the St2-only

state is rarely conserved across species (Atallah & Lott, 2018) for a given gene. It is possible that maternal deposition, where transcripts are dumped in the egg by the nurse cells, sometimes introduces developmental “noise” (the unnecessary deposition of RNA transcripts), and that the embryo compensates by degrading those transcripts.

Class 2: Stage 2 higher, stage 5 lower

These genes are represented at both stages, with lower levels at the later stage, and have maternally-supplied transcripts that are partially degraded by stage 5. (It is impossible to rule out that some of these genes may be zygotically transcribed, with the transcription compensating for at least some of the degradation.) We found 728 genes in this category with significant differential expression. We found the identified *D. melanogaster* orthologs to be enriched in the GO terms “mitochondrion”, “mitochondrial matrix” and “metabolic pathways”, although there was no significant enrichment for any specific pathway (Table S2). Furthermore, there was no significant enrichment of any GO terms in the Biological Process (BP), Cellular Component (CC) or Molecular Function (MF) categories. As with Class 1, genes with partially degraded transcripts are not limited to a specific subset of categories,

Class 3: Stage 2 lower, stage 5 higher

This class of genes include those that are represented at both stages, with levels that increase during development, and are believed to be both maternally supplied and zygotically transcribed. We found 1,364 genes in this category with significant differential expression between the two stages. They were enriched in products that play roles in protein and mRNA binding (Figure 3). Cellular components are localized to the plasma membrane and cytoplasm.

They are also enriched in components of the Hippo and Notch signaling pathways, and have functions in RNA transport and ubiquitin-mediated proteolysis. Many of these genes encode protein kinases. Although genes in the Wnt signaling pathway were not significantly enriched, enrichment was seen for the InterPro (Hunter et al., 2009) term “armadillo-type fold”. Tandem armadillo repeats are found in many proteins involved in Wnt signaling. Signaling pathways play key roles during cellularization (stage 5), and segment polarity genes include components of Wnt signaling. The provision of these components through two mechanisms (maternal deposition and zygotic transcription) could be a form of redundancy to increase developmental robustness (Mestek Boukhibar & Barkoulas, 2016).

Class 4: Stage 5-only

We found 880 zygotic-only genes (represented only at St5). They encompass a range of important players in early development, and include genes involved in patterning, development of the central nervous system (including the brain) and several other ontogenetic processes (Figure 4A). Since St5 is the period when cellularization occurs, it is not surprising that more genes map to the GO term “integral component of membrane” than to any other cellular component (Figure 4B). They include genes encoding proteins localized to the plasma cell membrane and organelle membranes. Many of the gene products are also part of the extracellular matrix.

Not surprisingly, transcription factors, which activate patterning cascades and jump-start development, are by far the most strongly overrepresented molecular function class of St5-only genes, both in terms of the number of genes mapping to this category and the negative

logarithm of the q-value (Figure 4C). Other zygotic-only genes have serine-type endopeptidase activity (GO term 0004252 in in Figure 4C); these include *masquerade*, which has a known role axonal development, and *modular serine protease* (modSP), which is involved in the immune system. St5-only genes also include heme-binding enzymes with oxidoreductive activity, including an array of cytochrome P450 enzymes (e.g. *phantom*, *disembodied*, *spookier* and many others).

Some gain, and widespread loss, of gene expression in the *D. grimshawi* stage 5 embryo

The results of the differential expression analysis above were largely concordant with our previous findings for non-Hawaiian species (Atallah & Lott, 2018). We were curious, however, about any differences that might exist in gene representation in *D. grimshawi*, the first Hawaiian species with embryonic transcriptomes. When considering genes with *D. melanogaster* orthologs, we found more cases (101) of unique losses of stage 5 representation in *D. grimshawi* than in any other species (Figure 6).

Among the most interesting losses of St5 gene representation are the homeotic selector (Hox) genes, the downstream component of the anterior-posterior (AP) patterning pathway. As can be seen in Figure 7, multiple Hox genes are expressed by stage 5 in all species except *D. grimshawi*. We confirmed that components of the AP cascade upstream of the Hox genes, including the segment polarity, pair-rule and gap genes, are all expressed by stage 5 in *D. grimshawi* (Table S3).

There are also a total of 79 genes with unique gains of representation (Figure 6). We took a closer look at the 35 genes with transcript abundance three times our threshold (FPKM > 3).

Within this smaller set, 8 genes have named *D. melanogaster* orthologs, and are listed in Table S4. They include *Lim3*, a zinc finger transcription factor with a homeobox domain with a role in motor-neuron development. Another gene, *multiple wing hairs (mwh)*, a downstream component of the planar cell polarity pathway, is represented in at both St2 and St5 in *D. grimshawi*, but not in any other species we have examined.

Strong expression of unannotated genes in *D. grimshawi*

RNA-Seq bioinformatics methods that include the discovery of new isoforms, such as the Tuxedo suite, often uncover previously unannotated genes. In *D. grimshawi*, we find more unannotated transcripts with high stage 5 embryonic mRNA levels (FPKM > 3) than in any other species we have previously studied (Figure 7A). This finding could be partially due to the relatively poor annotation of the genome of this Hawaiian species (although the *D. grimshawi* genome annotation has improved markedly in recent years (Yang et al., 2018)). However, we were interested in knowing whether some of these transcripts could belong to putative taxonomically-restricted protein-coding genes.

A previous study in Drosophilidae (Heames et al., 2020) found unannotated genes in the family to be relatively short, coding for a median peptide length of only 81 amino acids. Using TransDecoder (Haas et al., 2013), we identified the longest complete open reading frame (ORF) (if any) in each of the 2,365 unannotated transcripts, and found 969 transcripts with an ORF of at least 50 amino acids. We used BLASTP (Altschul et al., 1990) to determine whether these putative peptides had homologs in the annotated genomes or embryonic transcriptomes of other *Drosophila* species. A total of 854 of the 969 ORFs had no identifiable orthologs. Of

these, we chose to focus on the 301 transcripts with an FPKM above our threshold of 1 at Stage 2, Stage 5, or both.

A subset of unannotated genes may have been generated de novo from non-coding sequence

Taxonomically restricted genes may have either originated from ancestral genes (e.g. through divergence beyond recognition from an ortholog, or through gene fission or fusion) or emerged de novo from non-coding sequence (Van Oss & Carvunis, 2019). Examples of de novo gene evolution, which at one time was dismissed as highly improbable, have fascinated researchers in recent years (Klasberg et al., 2018; Neme & Tautz, 2016; Zhao et al., 2014). Researchers frequently use TBLASTN to identify putative intergenic or intronic regions in other species that orphan genes might have arisen from (Lu et al., 2017; Sun et al., 2015). We adopted a similar approach, aided by the recent publication (B. Y. Kim et al., 2021) of new *Drosophila* genome assemblies (as yet unannotated), generated through Oxford Nanopore long-read sequencing, that included two additional Hawaiian picture-wing species (*Drosophila murphyi* and *Drosophila sproati*) along with 4 *Scaptomyza* species (*Scaptomyza graminum*, *Scaptomyza hsui*, *Scaptomyza montana* and *Scaptomyza pallida*.)

Of the 114 transcripts with TBLASTN hits in one or more of the non-Hawaiian annotated genomes, the vast majority of the top hits (95, or 85%) are in intergenic or intronic regions in all species, suggesting that they may have arisen de novo from non-coding sequence (Figure 9A). Interestingly (Figure 8B), 97% of the 301 transcripts (291) had TBLASTN hits in either *D. murphyi* or *D. sproati* (the other two picture-wing Hawaiian *Drosophila* species). Only 47% had hits in the

genomes of one of the four species in the *Scaptomyza* lineage, and 35% in the *Drosophila* subgenus (*D. virilis* or *D. mojavensis*).

Of the transcripts with noncoding TBLASTN hits in annotated non-Hawaiian species genomes, 67 out of 95 (71%) matched regions that contain an ORF in at least one of the fly genomes (Figure 9A), possibly lending support to an ORF-first model of de novo gene evolution.

Alignments with CLUSTALW showed that in most cases both the start and stop codons were conserved in at least one of the other picture-wing *Drosophila* species (Figure 9B). The same was true far less frequently in *Scaptomyza* and the *Drosophila* subgenus species. We also found cases, however, where a start or stop codon may have been generated from non-coding sequence through a single-nucleotide substitution (Figure 9C). Our results suggest that a subset of unannotated genes with embryonic mRNA representation may have been generated de novo from non-coding sequence.

DISCUSSION

Island endemics are excellent systems for exploring the evolution of novelty. Most famously, the flora and fauna of the Galapagos islands were critical to Charles Darwin's (Darwin, 1845) development of the theory of natural selection. More recently, the evolution of *Anolis* lizards on the Greater and Lesser Antilles has been analyzed rigorously (Corbett-Detig et al., 2020; Mahler et al., 2010, 2016). Islands such as the Hawaiian archipelago, which arose within the past five million years through volcanic eruptions, presented pristine, untouched environments for early colonists with untapped ecological opportunities (Whittaker et al., 2017). Evolution on

these types of islands proceeds far more rapidly than on the mainland (Losos & Ricklefs, 2009), leading to diversity which often exceeds those of continental species.

As close relatives of the model organism *D. melanogaster*, the Hawaiian *Drosophila* are a particularly valuable model for studying the evolutionary genetics of island radiations. Research on this clade has slowed, however, in recent decades (O’Grady & DeSalle, 2018). We present here the first transcriptomic comparison in a Hawaiian fly of two iconic stages in embryogenesis that we know to be highly conserved across Drosophilidae (Kuntz & Eisen, 2014). We find that while *D. grimshawi*’s early embryonic transcriptome at these stages is similar in many ways to other non-Hawaiian species, there are distinct differences, with numerous losses of orthologous gene representation. Most notably, while the Hox genes are represented by stage 5 in every other species we have examined, we find no evidence of this in *D. grimshawi*. Future studies will be necessary to determine when homeotic selector genes are activated in this species, and whether the apparent delay in Hox activation is also seen in other Hawaiian flies.

Genes with zygotic (i.e. stage 5) expression in *D. grimshawi*, but not in other species (Table S4), are excellent candidates for functional knock-out experiments. We are currently developing protocols for targeted mutagenesis in *D. grimshawi* using the CRISPR-Cas9 system (Doudna & Charpentier, 2014; Jinek et al., 2012). The distinct advantage of examining the early embryo is that the effects of mutations can be assessed at the beginning of development. While in the long-term we would like to examine the activity and functions of genes active in larval and pupal discs of structures such as wings and legs, which have diverged markedly in Hawaiian species (Edwards et al., 2007; Stark & O’Grady, 2010), these goals will be more challenging to achieve, because the role of a gene early in development may mask its function at a later stage

(e.g. if the gene is embryonically lethal, it is more difficult to determine its later role in imaginal discs). Such research may require precise editing of specific enhancers.

Gains and losses of expression of genes with orthologs in other species, including the model fly *D. melanogaster*, could be examples of co-option of a conserved genetic toolkit (Carroll, 2008), although the functional experiments described above will be necessary for investigating this possibility. While the re-use of existing genes is thought to play an important role in developmental evolution (Stern, 2011), attention has turned more recently to the role of taxonomically restricted genes (Johnson, 2018; McLysaght & Guerzoni, 2015; Van Oss & Carvunis, 2019). Much of the work investigating de novo gene evolution has focused on transcripts expressed in the testis (Lange et al., 2021; Witt et al., 2019; Zhao et al., 2014), where evolutionary turnover occurs rapidly (Jagadeeshan & Singh, 2005; Mohammed et al., 2014), with fewer studies looking at genes expressed in relatively slowly-evolving systems such as embryogenesis. In this study, we were interested in the fact that more unannotated genes were expressed in *D. grimshawi* at St5 than in any of the 14 other species we had previously examined. While we had speculated (Atallah & Lott, 2018) that unannotated genes might be taxonomically restricted, we had not examined this possibility further, and it is possible that many unannotated genes are simply a result of poor genome annotation. In this study, we identified a total of 95 embryonic *D. grimshawi* transcripts with open reading frames that had homologous sequence matches in intergenic or intronic regions in a non-Hawaiian species (Figure 9A), suggesting that they may have originated de novo, either in the Hawaiian *Drosophila* lineage or the ancestor of Hawaiian *Drosophila* and *Scaptomyza*.

Two models exist to explain the emergence of de novo protein-coding genes (McLysaght & Guerzoni, 2015). In the “ORF-first” model, a pre-existing ORF is expressed through regulatory element evolution. In a second model, a non-coding RNA acquires an ORF (Xie et al., 2012). Analyzing embryonic transcriptomes from other Hawaiian *Drosophila* and *Scaptomyza* species, where we often found ORFs in regions that were homologous to the unannotated expressed sequences in *D. grimshawi*, will be necessary to distinguish between these models. As with co-opted genes, targeted mutagenesis of putative de novo genes will be important for determining whether they have acquired novel functions in the early embryo.

ACKNOWLEDGEMENTS

We would like to thank Delmy Urbina, Jacob Michalak and Bianca Canizares for assistance with the experiments. Funding for this work was provided to Joel Atallah from the Louisiana Board of Regents (grant number LEQSF(2017-20)-RD-A-26).

DATA AVAILABILITY STATEMENT

REFERENCES

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- Atallah, J., & Lott, S. E. (2018). Evolution of maternal and zygotic mRNA complements in the early *Drosophila* embryo. *PLOS Genetics*, 14(12), e1007838. <https://doi.org/10.1371/journal.pgen.1007838>

- Bownes, M. (1975). A photographic study of development in the living embryo of *Drosophila melanogaster*. *Journal of Embryology and Experimental Morphology*, 33(3), 789–801.
- Carroll, S. B. (2008). Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell*, 134(1), 25–36. <https://doi.org/10.1016/j.cell.2008.06.030>
- Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., Kaufman, T. C., Kellis, M., Gelbart, W., Iyer, V. N., Pollard, D. A., Sackton, T. B., Larracuenta, A. M., Singh, N. D., Abad, J. P., Abt, D. N., Adryan, B., Aguade, M., Akashi, H., ... MacCallum, I. (2007). Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, 450(7167), 203–218. <https://doi.org/10.1038/nature06341>
- Combs, P. A., & Eisen, M. B. (2013). Sequencing mRNA from Cryo-Sliced *Drosophila* Embryos to Determine Genome-Wide Spatial Patterns of Gene Expression. *PLOS ONE*, 8(8), e71820. <https://doi.org/10.1371/journal.pone.0071820>
- Corbett-Detig, R. B., Russell, S. L., Nielsen, R., & Losos, J. (2020). Phenotypic Convergence Is Not Mirrored at the Protein Level in a Lizard Adaptive Radiation. *Molecular Biology and Evolution*, 37(6), 1604–1614. <https://doi.org/10.1093/molbev/msaa028>
- Craddock, E. M., Kambyzellis, M. P., Franchi, L., Francisco, P., Grey, M., Hutchinson, A., Nanhoo, S., & Antar, S. (2018). Ultrastructural variation and adaptive evolution of the ovipositor in the endemic Hawaiian *Drosophilidae*. *Journal of Morphology*, 279(12), 1725–1752. <https://doi.org/10.1002/jmor.20884>
- Darwin, C. (1845). *Journal of Researches Into the Natural History and Geology of the Countries Visited During the Voyage of H.M.S. Beagle Round the World*. John Murray.
- Doudna, J. A., & Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. *Science*, 346(6213). <https://doi.org/10.1126/science.1258096>

- Edwards, K. A., Doescher, L. T., Kaneshiro, K. Y., & Yamamoto, D. (2007). A Database of Wing Diversity in the Hawaiian *Drosophila*. *PLOS ONE*, 2(5), e487. <https://doi.org/10.1371/journal.pone.0000487>
- Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., ... Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8(8), 1494–1512. <https://doi.org/10.1038/nprot.2013.084>
- Heames, B., Schmitz, J., & Bornberg-Bauer, E. (2020). A Continuum of Evolving De Novo Genes Drives Protein-Coding Novelty in *Drosophila*. *Journal of Molecular Evolution*, 88(4), 382–398. <https://doi.org/10.1007/s00239-020-09939-z>
- Hedges, S. B., Marin, J., Suleski, M., Paymer, M., & Kumar, S. (2015). Tree of Life Reveals Clock-Like Speciation and Diversification. *Molecular Biology and Evolution*, 32(4), 835–845. <https://doi.org/10.1093/molbev/msv037>
- Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009a). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1), 44–57. <https://doi.org/10.1038/nprot.2008.211>
- Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009b). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1), 1–13. <https://doi.org/10.1093/nar/gkn923>
- Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., ... Yeats, C. (2009). InterPro: The integrative protein signature database. *Nucleic Acids Research*, 37(suppl_1), D211–D215. <https://doi.org/10.1093/nar/gkn785>

- Jagadeeshan, S., & Singh, R. S. (2005). Rapidly Evolving Genes of *Drosophila*: Differing Levels of Selective Pressure in Testis, Ovary, and Head Tissues Between Sibling Species. *Molecular Biology and Evolution*, 22(9), 1793–1801. <https://doi.org/10.1093/molbev/msi175>
- Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*, 337(6096), 816–821. <https://doi.org/10.1126/science.1225829>
- Johnson, B. R. (2018). Taxonomically Restricted Genes Are Fundamental to Biology and Evolution. *Frontiers in Genetics*, 9. <https://doi.org/10.3389/fgene.2018.00407>
- Kambysellis, M. P., & Heed, W. B. (1971). Studies of Oogenesis in Natural Populations of *Drosophilidae*. I. Relation of Ovarian Development and Ecological Habitats of the Hawaiian Species. *The American Naturalist*, 105(941), 31–49.
- Kambysellis, M. P., Starmer, T., Smathers, G., & Heed, W. B. (1980). Studies of Oogenesis in Natural Populations of *Drosophilidae*. II. Significance of Microclimatic Changes on Oogenesis of *Drosophila mimica*. *The American Naturalist*, 115(1), 67–91.
- Kang, L., Settlage, R., McMahon, W., Michalak, K., Tae, H., Garner, H. R., Stacy, E. A., Price, D. K., & Michalak, P. (2016). Genomic Signatures of Speciation in Sympatric and Allopatric Hawaiian Picture-Winged *Drosophila*. *Genome Biology and Evolution*, 8(5), 1482–1488. <https://doi.org/10.1093/gbe/evw095>
- Katoh, T., Izumitani, H. F., Yamashita, S., & Watada, M. (2017). Multiple origins of Hawaiian drosophilids: Phylogeography of *Scaptomyza* Hardy (Diptera: *Drosophilidae*). *Entomological Science*, 20(1), 33–44. <https://doi.org/10.1111/ens.12222>
- Kim, B. Y., Wang, J. R., Miller, D. E., Barmina, O., Delaney, E., Thompson, A., Comeault, A. A., Peede, D., D’Agostino, E. R., Pelaez, J., Aguilar, J. M., Haji, D., Matsunaga, T., Armstrong, E. E., Zych, M., Ogawa, Y., Stamenković-Radak, M., Jelić, M., Veselinović, M. S., ... Petrov, D. A. (2021). Highly

- contiguous assemblies of 101 drosophilid genomes. *ELife*, 10, e66405.
<https://doi.org/10.7554/eLife.66405>
- Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14, R36. <https://doi.org/10.1186/gb-2013-14-4-r36>
- Klasberg, S., Bitard-Feildel, T., Callebaut, I., & Bornberg-Bauer, E. (2018). Origins and structural properties of novel and de novo protein domains during insect evolution. *The FEBS Journal*, 285(14), 2605–2625. <https://doi.org/10.1111/febs.14504>
- Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution*, 34(7), 1812–1819. <https://doi.org/10.1093/molbev/msx116>
- Kuntz, S. G., & Eisen, M. B. (2014). Drosophila Embryogenesis Scales Uniformly across Temperature in Developmentally Diverse Species. *PLOS Genetics*, 10(4), e1004293. <https://doi.org/10.1371/journal.pgen.1004293>
- Lange, A., Patel, P. H., Heames, B., Damry, A. M., Saenger, T., Jackson, C. J., Findlay, G. D., & Bornberg-Bauer, E. (2021). Structural and functional characterization of a putative de novo gene in Drosophila. *Nature Communications*, 12(1), 1667. <https://doi.org/10.1038/s41467-021-21667-6>
- Lapoint, R. T., Magnacca, K. N., & O’Grady, P. M. (2009). Review of the spoon tarsus subgroup of Hawaiian Drosophila (Drosophilidae: Diptera), with a description of one new species. *Zootaxa*, 2003, 53–68.
- Lapoint, R. T., Magnacca, K. N., & O’Grady, P. M. (2014). Phylogenetics of the Antopocerus - Modified Tarsus Clade of Hawaiian Drosophila: Diversification across the Hawaiian Islands. *PLOS ONE*, 9(11), e113227. <https://doi.org/10.1371/journal.pone.0113227>

- Lapoint, R. T., O'Grady, P. M., & Whiteman, N. K. (2013). Diversification and dispersal of the Hawaiian Drosophilidae: The evolution of Scaptomyza. *Molecular Phylogenetics and Evolution*, 69(1), 95–108. <https://doi.org/10.1016/j.ympev.2013.04.032>
- Larkin, A., Marygold, S. J., Antonazzo, G., Attrill, H., dos Santos, G., Garapati, P. V., Goodman, J. L., Gramates, L. S., Millburn, G., Strelets, V. B., Tabone, C. J., Thurmond, J., & FlyBase Consortium. (2021). FlyBase: Updates to the Drosophila melanogaster knowledge base. *Nucleic Acids Research*, 49(D1), D899–D907. <https://doi.org/10.1093/nar/gkaa1026>
- Losos, J. B., & Ricklefs, R. E. (2009). Adaptation and diversification on islands. *Nature*, 457(7231), 830–836. <https://doi.org/10.1038/nature07893>
- Lott, S. E., Villalta, J. E., Zhou, Q., Bachtrog, D., & Eisen, M. B. (2014). Sex-Specific Embryonic Gene Expression in Species with Newly Evolved Sex Chromosomes. *PLoS Genet*, 10(2), e1004159. <https://doi.org/10.1371/journal.pgen.1004159>
- Lu, T.-C., Leu, J.-Y., & Lin, W.-C. (2017). A Comprehensive Analysis of Transcript-Supported De Novo Genes in Saccharomyces sensu stricto Yeasts. *Molecular Biology and Evolution*, 34(11), 2823–2838. <https://doi.org/10.1093/molbev/msx210>
- Magnacca, K. N., & Grady, P. M. (2009). *Revision of the Modified Mouthparts Species Group of Hawaiian Drosophila (Diptera: Drosophilidae): The Ceratostoma, Freycinetiae, Semifuscata, and Setiger Subgroups, and Unplaced Species* (Vol. 130). Univ of California Press.
- Magnacca, K. N., & Price, D. K. (2015). Rapid adaptive radiation and host plant conservation in the Hawaiian picture wing Drosophila (Diptera: Drosophilidae). *Molecular Phylogenetics and Evolution*, 92, 226–242. <https://doi.org/10.1016/j.ympev.2015.06.014>
- Mahler, D. L., Lambert, S. M., Geneva, A. J., Ng, J., Hedges, S. B., Losos, J. B., & Glor, R. E. (2016). Discovery of a Giant Chameleon-Like Lizard (Anolis) on Hispaniola and Its Significance to

- Understanding Replicated Adaptive Radiations. *The American Naturalist*, 188(3), 357–364.
<https://doi.org/10.1086/687566>
- Mahler, D. L., Revell, L. J., Glor, R. E., & Losos, J. B. (2010). Ecological Opportunity and the Rate of Morphological Evolution in the Diversification of Greater Antillean Anoles. *Evolution*, 64(9), 2731–2745. <https://doi.org/10.1111/j.1558-5646.2010.01026.x>
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, 17(1), 10–12. <https://doi.org/10.14806/ej.17.1.200>
- McLysaght, A., & Guerzoni, D. (2015). New genes from non-coding sequence: The role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678), 20140332.
<https://doi.org/10.1098/rstb.2014.0332>
- Mestek Boukhibar, L., & Barkoulas, M. (2016). The developmental genetics of biological robustness. *Annals of Botany*, 117(5), 699–707. <https://doi.org/10.1093/aob/mcv128>
- Mohammed, J., Bortolamiol-Becet, D., Flynt, A. S., Gronau, I., Siepel, A., & Lai, E. C. (2014). Adaptive evolution of testis-specific, recently evolved, clustered miRNAs in *Drosophila*. *RNA*, 20(8), 1195–1209. <https://doi.org/10.1261/rna.044644.114>
- Montgomery, S. L. (1975). Comparative breeding site ecology and the adaptive radiation of picture-winged *Drosophila* (Diptera: Drosophilidae) in Hawaii. *Proceedings, Hawaiian Entomological Society*, 22(1).
- Neme, R., & Tautz, D. (2016). Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *ELife*, 5, e09977.
<https://doi.org/10.7554/eLife.09977>
- O’Grady, P., & DeSalle, R. (2008). Out of Hawaii: The origin and biogeography of the genus *Scaptomyza* (Diptera: Drosophilidae). *Biology Letters*, 4(2), 195–199. <https://doi.org/10.1098/rsbl.2007.0575>

- O'Grady, P., & DeSalle, R. (2018). Hawaiian *Drosophila* as an Evolutionary Model Clade: Days of Future Past. *BioEssays*, 40(5), 1700246. <https://doi.org/10.1002/bies.201700246>
- Ometto, L., Cestaro, A., Ramasamy, S., Grassi, A., Revadi, S., Siozios, S., Moretto, M., Fontana, P., Varotto, C., Pisani, D., Dekker, T., Wrobel, N., Viola, R., Pertot, I., Cavalieri, D., Blaxter, M., Anfora, G., & Rota-Stabelli, O. (2013). Linking Genomics and Ecology to Investigate the Complex Evolution of an Invasive *Drosophila* Pest. *Genome Biology and Evolution*, 5(4), 745–757. <https://doi.org/10.1093/gbe/evt034>
- Paris, M., Villalta, J. E., Eisen, M. B., & Lott, S. E. (2015). Sex Bias and Maternal Contribution to Gene Expression Divergence in *Drosophila* Blastoderm Embryos. *PLOS Genet*, 11(10), e1005592. <https://doi.org/10.1371/journal.pgen.1005592>
- Sarikaya, D. P., Church, S. H., Lagomarsino, L. P., Magnacca, K. N., Montgomery, S. L., Price, D. K., Kaneshiro, K. Y., & Extavour, C. G. (2019). Reproductive Capacity Evolves in Response to Ecology through Common Changes in Cell Number in Hawaiian *Drosophila*. *Current Biology*, 29(11), 1877-1884.e6. <https://doi.org/10.1016/j.cub.2019.04.063>
- Stark, J. B., & O'Grady, P. M. (2010). Morphological variation in the forelegs of the Hawaiian *Drosophilidae*. I. The AMC clade. *Journal of Morphology*, 271(1), 86–103. <https://doi.org/10.1002/jmor.10783>
- Stern, D. L. (2011). *Evolution, development, & the predictable genome*. Roberts and Co. Publishers. <http://agris.fao.org/agris-search/search.do?recordID=US201300151702>
- Sun, W., Zhao, X.-W., & Zhang, Z. (2015). Identification and evolution of the orphan genes in the domestic silkworm, *Bombyx mori*. *FEBS Letters*, 589(19, Part B), 2731–2738. <https://doi.org/10.1016/j.febslet.2015.08.008>
- Tadros, W., & Lipshitz, H. D. (2009). The maternal-to-zygotic transition: A play in two acts. *Development*, 136(18), 3033–3042. <https://doi.org/10.1242/dev.033183>

- Thompson, J. D., Gibson, Toby. J., & Higgins, D. G. (2003). Multiple Sequence Alignment Using ClustalW and ClustalX. *Current Protocols in Bioinformatics*, 00(1), 2.3.1-2.3.22.
<https://doi.org/10.1002/0471250953.bi0203s00>
- Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, 31(1), 46–53. <https://doi.org/10.1038/nbt.2450>
- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., & Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7(3), 562–578.
<https://doi.org/10.1038/nprot.2012.016>
- Van Oss, S. B., & Carvunis, A.-R. (2019). De novo gene birth. *PLOS Genetics*, 15(5), e1008160.
<https://doi.org/10.1371/journal.pgen.1008160>
- Walter, W., Sánchez-Cabo, F., & Ricote, M. (2015). GOplot: An R package for visually combining expression data with functional analysis. *Bioinformatics*, 31(17), 2912–2914.
<https://doi.org/10.1093/bioinformatics/btv300>
- Wheeler, M. R., & Clayton, F. (1965). A new *Drosophila* culture technique. *Drosophila Information Service*, 40, 98.
- Whittaker, R. J., Fernández-Palacios, J. M., Matthews, T. J., Borregaard, M. K., & Triantis, K. A. (2017). Island biogeography: Taking the long view of nature’s laboratories. *Science*, 357(6354).
<https://doi.org/10.1126/science.aam8326>
- Witt, E., Benjamin, S., Svetec, N., & Zhao, L. (2019). Testis single-cell RNA-seq reveals the dynamics of de novo gene transcription and germline mutational bias in *Drosophila*. *ELife*, 8, e47138.
<https://doi.org/10.7554/eLife.47138>

- Xie, C., Zhang, Y. E., Chen, J.-Y., Liu, C.-J., Zhou, W.-Z., Li, Y., Zhang, M., Zhang, R., Wei, L., & Li, C.-Y. (2012). Hominoid-Specific De Novo Protein-Coding Genes Originating from Long Non-Coding RNAs. *PLOS Genetics*, 8(9), e1002942. <https://doi.org/10.1371/journal.pgen.1002942>
- Yang, H., Jaime, M., Polihronakis, M., Kanegawa, K., Markow, T., Kaneshiro, K., & Oliver, B. (2018). Re-annotation of eight *Drosophila* genomes. *Life Science Alliance*, 1(6). <https://doi.org/10.26508/lsa.201800156>
- Zhao, L., Saelao, P., Jones, C. D., & Begun, D. J. (2014). Origin and Spread of de Novo Genes in *Drosophila melanogaster* Populations. *Science*, 343(6172), 769–772. <https://doi.org/10.1126/science.1248286>

FIGURE LEGENDS

Figure 1. Hawaiian *Drosophila*. A) *Drosophila grimshawi*. B) A map of the Hawaiian islands, with arrows showing inferred colonization events. Kaua’i is the oldest of the larger islands and Hawai’i the youngest. Maui Nu’i consists of Moloka’i, Lana’i, Maui and Kaho’olawe. Most colonization events occurred from older (northern) to younger (southern) islands, although there were also cases of reverse colonization. C) A phylogeny of 20 Drosophilidae species. The Hawaiian *Drosophila* lineage is a sister-group to *Scaptomyza*. The ancestor of both clades was probably Hawaiian, but *Scaptomyza* dispersed globally, while the Hawaiian *Drosophila* remained confined to the archipelago. The phylogeny is based on multiple studies (Kato et al., 2017; Magnacca & Price, 2015; Ometto et al., 2013) and divergence times gleaned from TimeTree (Hedges et al., 2015; Kumar et al., 2017) were used in its construction.

Figure 2. The maternal-to-zygotic transition (MZT). The zygotic genome is silent in the early embryo, which relies on transcripts that were deposited by the mother and translated after egg

activation to jump-start development. Many of these transcripts are degraded. By stage 5, the embryonic transcriptome consists of both maternally deposited mRNAs that have not been degraded and zygotically transcribed mRNAs.

Figure 3. Gene Ontology (GO) analysis of genes that are represented at both stages but are higher at St5 (Class 3). The height of the bars in the inner circle represents the Benjamini-corrected enrichment P-value of the associated term, and the color represents the “z-score”, in this case the square root of the number of genes mapping to the term. The outer circle shows the log of the fold enrichment (St5/St2) of genes mapping to each term. GOplot (Walter et al., 2015) was used to generate the plot.

Figure 4. GO analysis of St5-only genes (Class 4). See the legend to Figure 3 for a description of the bars and circles on each panel. A) Biological Process GO terms. B) Cellular Component GO terms. C) Molecular Function GO terms.

Figure 5. Protein Domain enrichment of Class 4 genes, using InterPro.

Figure 6. Numbers of unique gains (green arrows) and losses (red arrows) of St2 and St5 gene representation in each species. The analysis was conducted with 15 *Drosophila* species, but only the 12 originally sequenced species are shown (*D. mauritiana*, *D. santomea* and *D. miranda* are not shown). *D. virilis* shows the largest number of gains at both St2 and St5, while *D. willistoni* shows the largest number of losses at stage 2, and *D. grimshawi* shows the largest number of losses at stage 5.

Figure 7. Loss of St5 representation of homeotic selector gene transcripts in *D. grimshawi*. St2 PPFM levels (close to 0 across all species) are shown in blue and St5 levels in red.

Figure 8. Unannotated genes in *D. grimshawi* show high expression levels, with a subset showing evidence of de novo evolution. A) More *D. grimshawi* unannotated genes are represented at high levels at stage 5 (FPKM > 3) than for any of the other species we have examined. B) Almost all identified ORFs in our candidate set of 301 *D. grimshawi* unannotated transcripts have TBLASTN hits in genomes of at least one other picture-wing Hawaiian *Drosophila* species, while slightly under half have hits in a *Scaptomyza* genome, and about a third have hits in *D. virilis* or *D. mojavensis*.

Figure 9. Putative de novo genes may have arisen from pre-existing ORFs or been generated from non-coding sequence. A) TBLASTN hits of possible *D. grimshawi* de novo transcripts (from the candidate set of 301) in non-Hawaiian species are overwhelmingly in intergenic or intronic regions, with a pre-existing ORF identified in the majority of cases. B) The start and stop codons of identified ORFs in *D. grimshawi* unannotated transcripts are highly conserved in other picture-wing species. Conservation is lower in *Scaptomyza* and the *Drosophila* subgenus. C) An example of a new ORF in a picture-wing Hawaiian *Drosophila* species. In the unannotated *D. grimshawi* transcript TCONS_41457, a start codon evolved from a single nucleotide change in the ancestor of the picture-wing Hawaiian *Drosophila* species. The alignment was generated using CLUSTALW.

SUPPORTING INFORMATION – TABLE LEGENDS

Table S1. Gene FPKM levels in previously analyzed species (Atallah & Lott, 2018) and *D. grimshawi*. Dmel = *Drosophila melanogaster*, Dsim = *Drosophila simulans*, Dsec = *Drosophila sechellia*, Dmau = *D. mauritiana*, Dyak = *Drosophila yakuba*, Dsan = *Drosophila santomea*, Dere

= *Drosophila erecta*, Dana = *Drosophila ananasse*, Dpse = *Drosophila pseudoobscura*, Dper = *Drosophila persimilis*, Dmir = *Drosophila miranda*, Dwil = *Drosophila willistoni*, Dmoj = *Drosophila mojavensis*, Dvir = *Drosophila virilis*, Dgri = *Drosophila grimshawi*.

Table S2. Gene Ontology analysis using DAVID (Huang et al., 2009a, 2009a).

Table S3. Segmentation gene FPKM levels at St2 and St5. *D. grimshawi* and 14 other (Atallah & Lott, 2018) species are shown.

Table S4. List of genes with named *D. melanogaster* one-to-one orthologs that are strongly represented at St5 in *D. grimshawi* (FPKM > 3) and not represented at that stage in 14 other *Drosophila* species.

Figure 1

Class I	Present at stage 2	Absent at stage 5	Maternally supplied transcripts that are entirely degraded
Class II	High at stage 2	Low at stage 5	Maternally supplied transcripts that are partially degraded
Class III	Low at stage 2	High at stage 5	Evidence of both maternal provision and zygotic transcription
Class IV	Absent at stage 2	Present at stage 5	No maternal provision, zygotic transcription

Figure 2

Figure 3

ID	Description
GO:0045944	positive regulation of transcription from RNA polymerase II promoter
GO:0007419	ventral cord development
GO:0007417	central nervous system development
GO:0006508	proteolysis
GO:0007366	periodic partitioning by pair rule gene
GO:0000122	negative regulation of transcription from RNA polymerase II promoter
GO:0007494	midgut development
GO:0010629	negative regulation of gene expression
GO:0001708	cell fate specification
GO:0007420	brain development

ID	Description
GO:0005576	extracellular region
GO:0005615	extracellular space
GO:0016021	integral component of membrane
GO:0031012	extracellular matrix
GO:0005578	proteinaceous extracellular matrix
GO:0005887	integral component of plasma membrane
GO:0005886	plasma membrane
GO:0009986	cell surface
GO:0031090	organelle membrane

ID	Description
GO:0003700	transcription factor activity, sequence-specific DNA binding
GO:0004252	serine-type endopeptidase activity
GO:0020037	heme binding
GO:0005506	iron ion binding
GO:0004497	monooxygenase activity
GO:0046982	protein heterodimerization activity
GO:0005549	odorant binding
GO:0004984	olfactory receptor activity
GO:0005272	sodium channel activity

Figure 4

ID	Description
PF00046	Homeobox domain
PF00089	Trypsin
PF07679	Immunoglobulin I-set domain
PF00010	Helix-loop-helix DNA-binding domain
PF00067	Cytochrome P450
PF13855	Leucine rich repeat
PF05306	Protein of unknown function (DUF733)
PF03022	Major royal jelly protein

Figure 5

Figure 6

Figure 7

Figure 8

A

Non-Hawaiian TBLASTN matches

■ Exonic ■ Intergenic or intronic - ORF ■ Intergenic or intronic - no ORF

B

Start and stop codon conservation

■ Start & stop codons conserved ■ Start only conserved
■ Stop only conserved ■ Neither stop nor start conserved

C

Start codon

Figure 9